EDA credit data analysis

- This case study aims to identify patterns which indicate if a client has difficulty paying their instalments which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc.
- This will ensure that the consumers capable of repaying the loan are not rejected. Identification of such applicants using EDA is the aim of this case study.

Business Understanding

- The loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history. Because of that, some consumers use it to their advantage by becoming a defaulter. Suppose you work for a consumer finance company which specialises in lending various types of loans to urban customers. You have to use EDA to analyse the patterns present in the data. This will ensure that the applicants capable of repaying the loan are not rejected.
-
- When the company receives a loan application, the company has to decide for loan approval based on the applicant's profile. Two types of risks are associated with the bank's decision:
- If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company
- If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company.
-
- The data given below contains the information about the loan application at the time of applying for the loan. It contains two types of scenarios:
- **The client with payment difficulties:** he/she had late payment more than X days on at least one of the first Y instalments of the loan in our sample,

- **All other cases:** All other cases when the payment is paid on time.
-
- When a client applies for a loan, there are four types of decisions that could be taken by the client/company):
- **Approved:** The Company has approved loan Application
- **Cancelled:** The client cancelled the application sometime during approval. Either the client changed her/his mind about the loan or in some cases due to a higher risk of the client, he received worse pricing which he did not want.
- **Refused:** The company had rejected the loan (because the client does not meet their requirements etc.).
- **Unused offer:** Loan has been cancelled by the client but at different stages of the process.
- In this case study, you will use EDA to understand how consumer attributes and loan attributes influence the tendency to default.

Business Objectives

- This case study aims to identify patterns which indicate if a client has difficulty paying their instalments which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc. This will ensure that the consumers capable of repaying the loan are not rejected. Identification of such applicants using EDA is the aim of this case study.

- In other words, the company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilise this knowledge for its portfolio and risk assessment.

- To develop your understanding of the domain, you are advised to independently research a little about risk analytics - understanding the types of variables and their significance should be enough.

Data Understanding

This dataset has 3 files as explained below:

- application_data.csv contains all the information of the client at the time of application. The data is about whether a client has payment difficulties.
- previous_application.csv contains information about the client’s previous loan data. It contains the data whether the previous application had been Approved, Cancelled, Refused or Unused offer.
- columns_description.csv is data dictionary which describes the meaning of the variables.

Note
application.csv

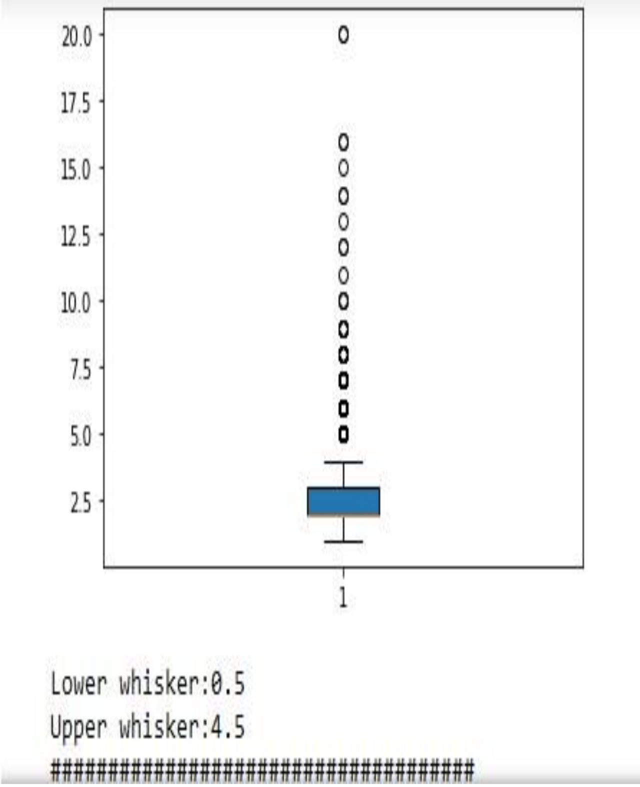
- Application data has 3,07,511 rows and 122 columns
- There are many null values in each columns
- Almost 49 columns out of 122 has more than 40% of null values

Note
Cleaning data approach

- Find out all columns which has more than 40% null values .
- Drop those columns.
- Columns which has less than 35% and more than %1 can be imputed to replace null values.
- Find out columns unique value which has <=3 values
- Convert those all values datatype to "object" so it will consider as

  Categorical data.
- Find key numerical columns and convert those columns to bin, so it will help to plot graph in visual range.

Outlier data frames

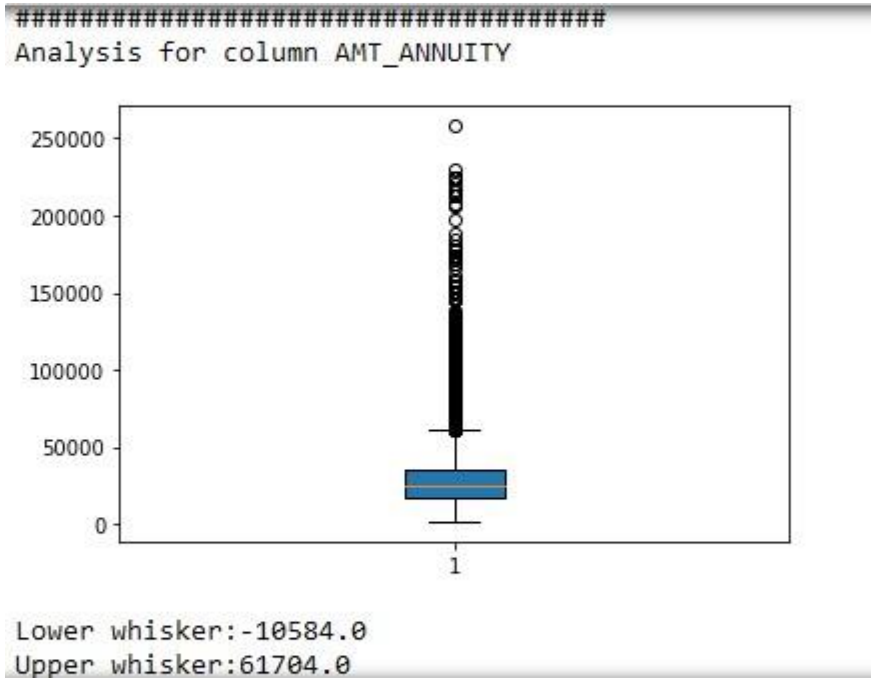


```
Lower whisker:0.5
Upper whisker:4.5
#####################################
```

- **CNT_FAM_MEMBERS column has outlier**
- **Applicant with more than 5 members are outliers**
- **All outliers can replace by mean**

Outlier data frames

- AMT_ANNUITY column
- As observed in box plot this column has outlier
- Applicants with AMT_ANNUITY above 61704 are outliers
- AMT_ANNUITY mean and 50% has difference so outliers can replace by median

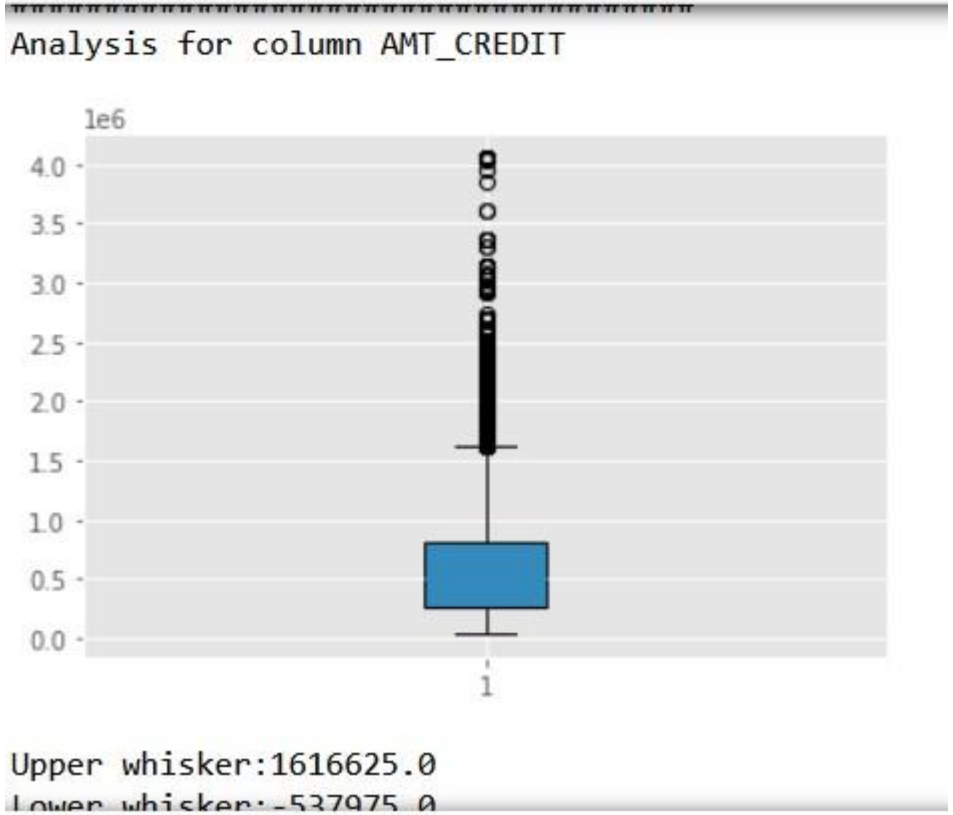


```
#####################################
Analysis for column AMT_ANNUITY
```

```
Lower whisker:-10584.0
Upper whisker:61704.0
```

Outlier data frames

- AMT_CREDIT column
- As observed in box plot this column has outlier
- Applicants with AMT_CREDIT above 1616625 are outliers
- This columns mean and 50% is almost close so we can replace all outliers with mean or median

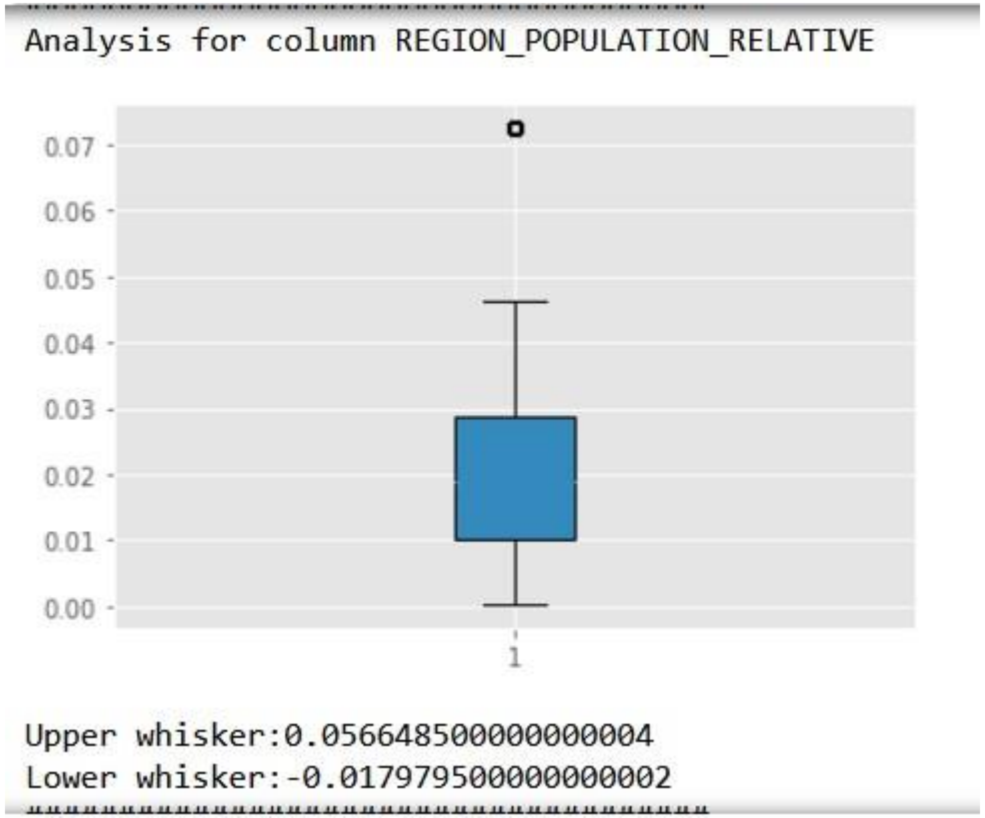


```
Analysis for column AMT_CREDIT
```

```
Upper whisker:1616625.0
Lower whisker:-537975.0
```

Outlier data frames

- REGION_POPULATION_RELATIVE column
- As observed in box plot this column has outlier
- Applicants with REGION_POPULATION_RELATIVE above 0.05 are outliers
- This columns mean and 50% is almost close so we can replace all outliers with mean or median



```
Analysis for column REGION_POPULATION_RELATIVE
```

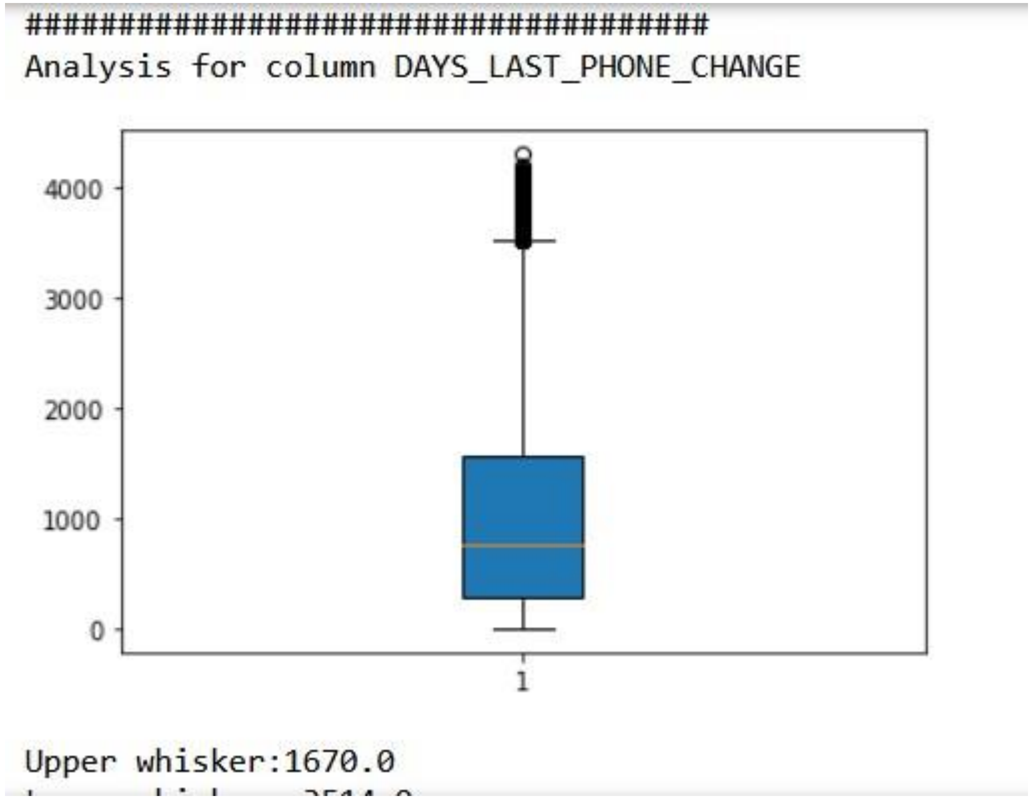
```
Upper whisker:0.056648500000000004
Lower whisker:-0.017979500000000002
```

Outlier data frames

- DAYS_LAST_PHONE_CHANGE

  column
- As observed in box plot this column has outlier
- Applicants with DAYS_LAST_PHONE_CHANGE

  above 1670 are outliers
- This columns mean and 50% has difference so we can replace all outliers with median



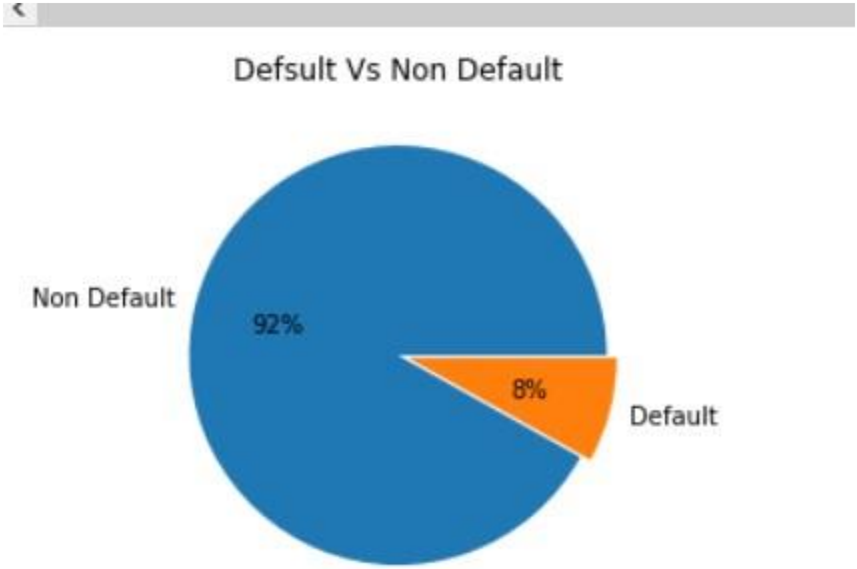
```
#####################################
Analysis for column DAYS_LAST_PHONE_CHANGE
```

```
Upper whisker:1670.0
```

Target analysis

- • We have imbalance in TARGET variable based on the % of observations

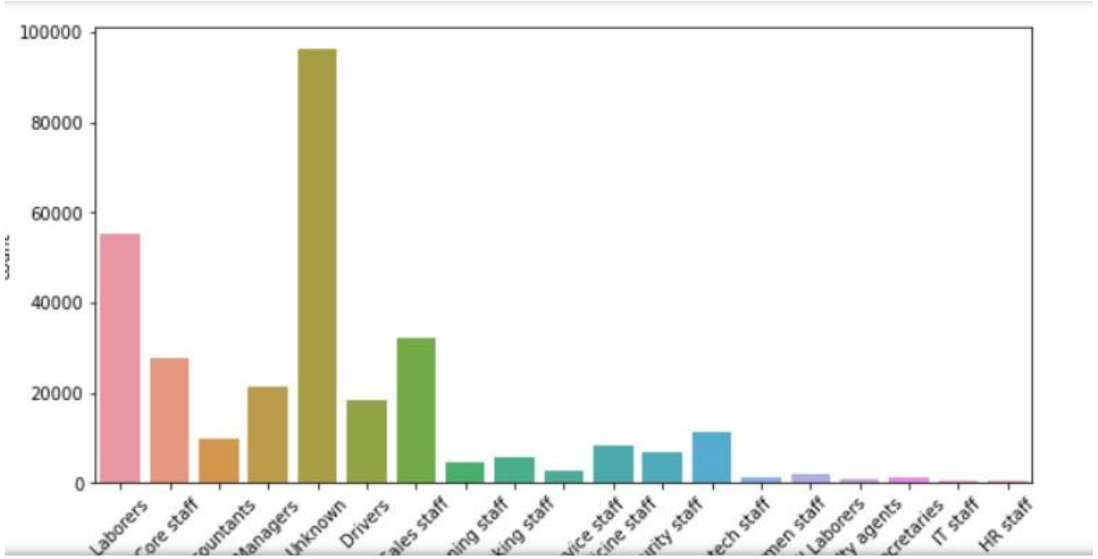
- as per counts there are 91.92% non default customer and 8.07% defaulter customer



Occupation type

- Most of loan’s applicants are belongs to unknows occupation type and 2nd highest applicants belongs to `Laborers` occupation

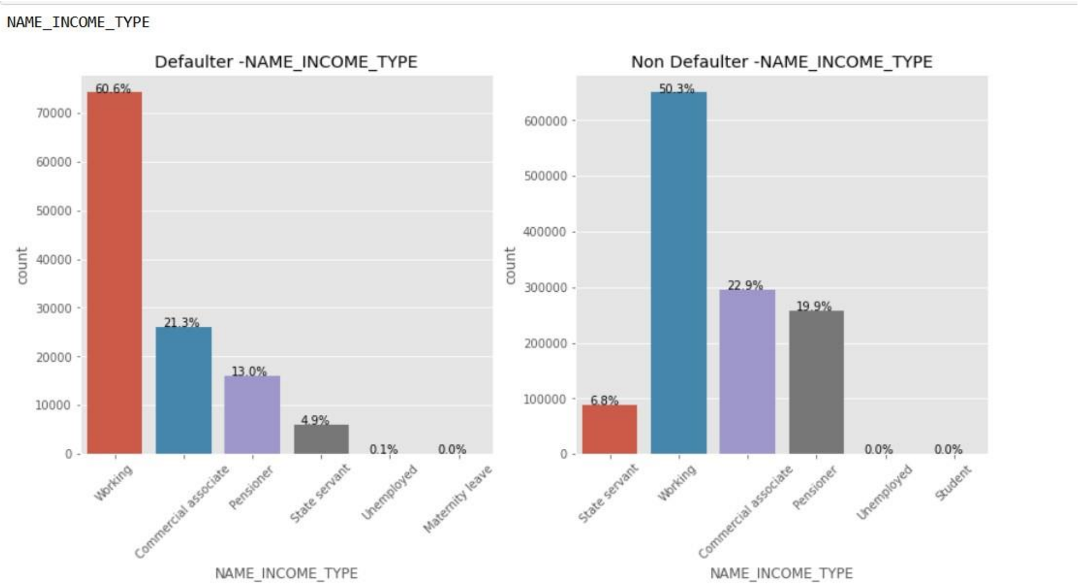
  Category.



Univariate analysis of categorical variables

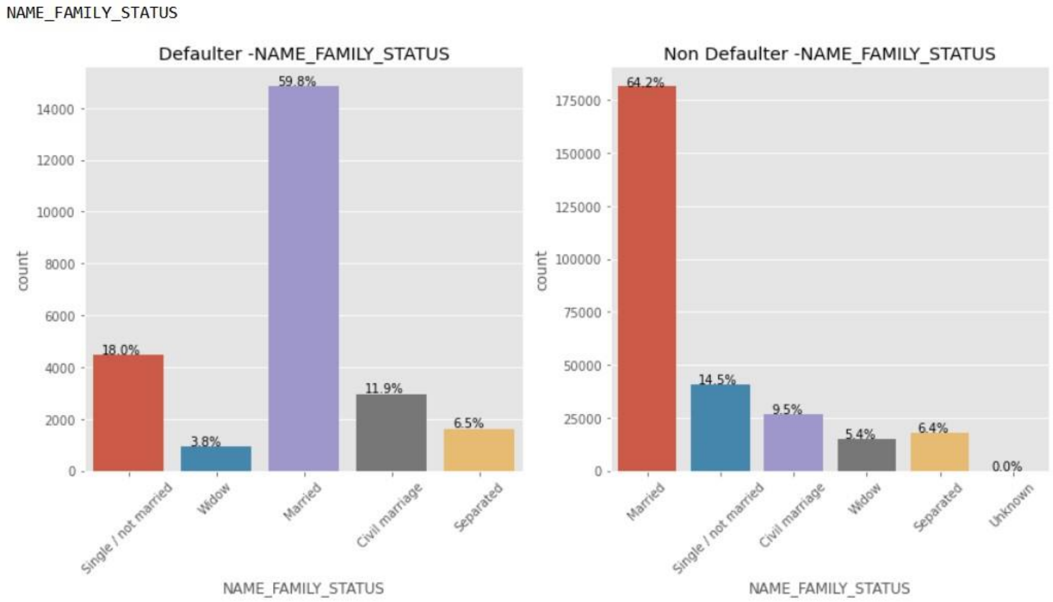
NAME_INCOME_TYPE

- Students don’t have any payment difficulties
- Working customers are better on time payment



Univariate analysis of categorical variable `Name_Family_Status`

- Clients who are married are 59.8% with defaulter customers and 64.2% with non defaulters
- Clients who are Widow are 3.8% with defaulter customers and 5.4% with non defaulters
- Clients who are Single/not married are 18.0% with defaulter customers and 14.5% with on non defaulters
- Clients who are married do payments on time better.


NAME_FAMILY_STATUS
Defaulter -NAME_FAMILY_STATUS
count
Single / not married
Widow
Married
Civil marriage
Separated
NAME_FAMILY_STATUS
Non Defaulter -NAME_FAMILY_STATUS
count
Married
Single / not married
Civil marriage
Widow
Separated
Unknown
NAME_FAMILY_STATUS